Available online at www.sciencedirect.com
ScienceDirect
Virology 366 (2007) 150–165
VIROLOGY
www.elsevier.com/locate/yviro

Natural genetic variations in bovine leukemia virus *envelope* gene: Possible effects of selection and escape☆

Xiangrong Zhao [a,*], Gertrude C. Buehring [b]

[a] *Graduate Program in Endocrinology, 3060 Valley Life Science Building, University of California, Berkeley, CA 94720-3140, USA*
[b] *Division of Infectious Diseases, School of Public Health, University of California, Berkeley, CA 94720-7360, USA*

Received 14 February 2007; returned to author for revision 5 March 2007; accepted 29 March 2007
Available online 11 May 2007

Abstract

Bovine leukemia virus (BLV) is an oncogenic virus widespread in cattle. It belongs to the genus *Deltaretrovirus* of the family *Retroviridae* along with human and simian T-lymphotropic viruses. Here we report the addition of 28 new sequences to the current literature of 16 full-length BLV *envelope* gene sequences. The phylogenetic clustering, genotyping, and geographic distribution of BLV *env* variations corresponded in most cases. Most natural variations are mapped to the surface of the proposed conformational models of BLV gp51 N-terminus and gp30 external domain, overlapping with or adjacent to immunogenic epitopes. Analyses for evidence of possible selection pressures suggest the BLV *env* is under stringent negative selection overall, while strong positive selection is indicated for immunogenic epitope G. Natural *env* deletions bounded by similar flanking sequences were observed in multiple isolates and would result in truncated signal peptides, missing gp51, and aberrant coding frames for other proteins.

Keywords: Bovine leukemia virus; *env*; Polymorphism; Phylogeny; Selection; Conformational model; Deletion

Introduction

Bovine leukemia virus (BLV) infects cattle worldwide and imposes severe financial impacts on the cattle industry. As a retrovirus, BLV has envelope (Env) proteins which contain functional domains that determine its infectivity (Johnston et al., 2002). BLV Env protein complex is composed of surface glycoprotein subunit (SU), which is anchored to the envelope via disulfide bonding (Johnston and Radke, 2000) with transmembrane glycoprotein subunit (TM). The newly synthesized BLV Env polyprotein precursor is cleaved into 3 pieces in the Golgi apparatus of the virus-producing cell: leader (signal) peptide (amino acid [aa] 1–33), SU (aa34–301), and TM (aa302–515). Oligomers of mature Env proteins are transported to the host cell surface membrane, where they are inserted. Viral particles bud at these insertion sites, acquiring an envelope with Env proteins in the process. In mature virions, SU peplomers recognize and bind to cell surface receptors, initiating a series of conformational changes that lead to the fusion of viral envelope and cellular surface membranes via TM oligomers. This is followed by the entry of the viral genome into the cell, digestion of the viral capsid, synthesis of the DNA provirus, integration into host genome (not necessarily occurring), and expression of viral genes.

The structure and functions of the BLV Env proteins have been extensively studied via delineation of conformational epitopes recognized by monoclonal antibodies (Bruck et al., 1982; Callebaut et al., 1991; Ban et al., 1992), and via identification by antipeptide antibodies (targeting linear epitopes) that block syncytium formation or neutralize the infectivity of pseudotype virus (Portetelle et al., 1989a; Callebaut et al., 1993). The N-terminal half of mature BLV gp51 SU plays an important role in viral infectivity and syncytium formation (Portetelle et al., 1989b; Callebaut et al., 1993), and probably contains the receptor-binding domain (Johnston et al., 2002). Also located there are the

☆ GenBank Accession Nos.: Nucleotide sequence data from this article have been deposited with the GenBank Data Libraries under Accession No. EF065635–EF065662.

* Corresponding author. Mailing address: 3060 Valley Life Science Building, University of California, Berkeley, CA 94720-3140, USA. Fax: +1 510 642 6350.
E-mail addresses: xzhao@berkeley.edu (X. Zhao), buehring@berkeley.edu (G.C. Buehring).

0042-6822/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2007.03.058

conformational epitopes F, G, and H (Figs. 1 and 2) (Bruck et al., 1982), whose recognition by specific monoclonal antibodies depends on disulfide bonding (Portetelle et al., 1989b) and glycosylation (Bruck et al., 1984). Monoclonal antibodies specific for the F, G, and H epitopes bound differentially to BLV Env proteins with slightly different amino acid sequences, leading to the identification of amino acids potentially affecting SU conformation (Portetelle et al., 1989b; Mamoun et al., 1990). The C-terminal half of BLV gp51 SU (Figs. 1 and 2) contains linear epitopes A, B, D, and E, mapped with antipeptide antibodies. This segment of SU also interacts with TM, probably via a disulfide bond (Johnston and Radke, 2000) formed between a conserved CXXC (X refers to any amino acid; Fig. 2) motif in SU (Gallaher et al., 1995) and a conserved CX_6CC motif in TM (Rice et al., 1984; Fig. 2).

The BLV gp30 TM (Figs. 1 and 2; Table 2) contains an extracellular domain (N-terminal) encompassing the hydrophobic fusion peptide (Gatot et al., 1998), a 4-3 heptad repeat (leucine zipper) (Gatot et al., 1998), and the GD21 peptide (Hadlock et al., 1995; Dube et al., 1997); a middle transmembrane domain hosting a stretch of 20 uncharged residues (aa438–457) spanning the host cell membrane before budding (Rice et al., 1984); and a cytoplasmic domain (C-terminal) including PXXP motifs (Cantor, 1996; Reichert et al., 2001) interspersed with YXXL immunoreceptor tyrosine activation motifs (ITAMs) (Inabe et al., 1999; Novakovic et al., 2004).

In contrast to many other retroviruses, e.g. human immunodeficiency virus-1 (HIV-1) (Liu et al., 2006), analyses of the BLV *env* gene sequences of isolates with different geographic origins demonstrated a high sequence conservation (Camargos et al., 2002, 2006). However, alterations in Env protein sequence resulting from non-synonymous nucleotide substitutions, although not frequent, could affect BLV infectivity and/or pathogenicity (Willems et al., 1995, 2000; Fechner et al., 1997).

BLV belongs to the *Deltaretrovirus* genus which also includes primate (human/simian) T-cell lymphotropic viruses (PTLVs): HTLV-1, -2 and -3, STLV-1, -2 and -3. Phylogenetically, BLV and PTLVs share a common ancestor. Studies on the BLV genetic diversity and its implications are not only of interest for veterinary medicine (e.g. pathogenicity, diagnosis, and vaccine development), but also for the possibility of viral transmission to other species including human (Buehring et al., 2003). So far there have been 16 full-length BLV *env* sequences deposited in sequence databanks (e.g. GenBank) by multiple research groups over 23 years. Here we report 28 newly determined full-length BLV *env* sequences, from field isolates collected from 4 different continents. Analyses for phylogenetic clustering, genotyping, and evidence of possible selection across as well as within clusters were conducted for a sample pool containing all those 44 full-length BLV *env* sequences available to date. Homology modeling for BLV gp51 N-terminus and gp30 external domain provided a hint of the possible spatial distribution of sequence variations. Natural extensive deletions in gp51 coding sequence, bound by similar flanking sequences, were observed in some samples and the possible biological effects of this are discussed.

Results and discussion

Variations were detected in BLV env gene and deduced Env protein sequences

All cattle from which BLV isolates were obtained were determined to be BLV-positive by serological tests (enzyme-linked immunosorbent assay, immunoblotting, and/or agar gel immunodiffusion assay) performed in the country of origin (data not shown). Twenty-eight full-length BLV *env* sequences were obtained by nested PCR amplification, followed by

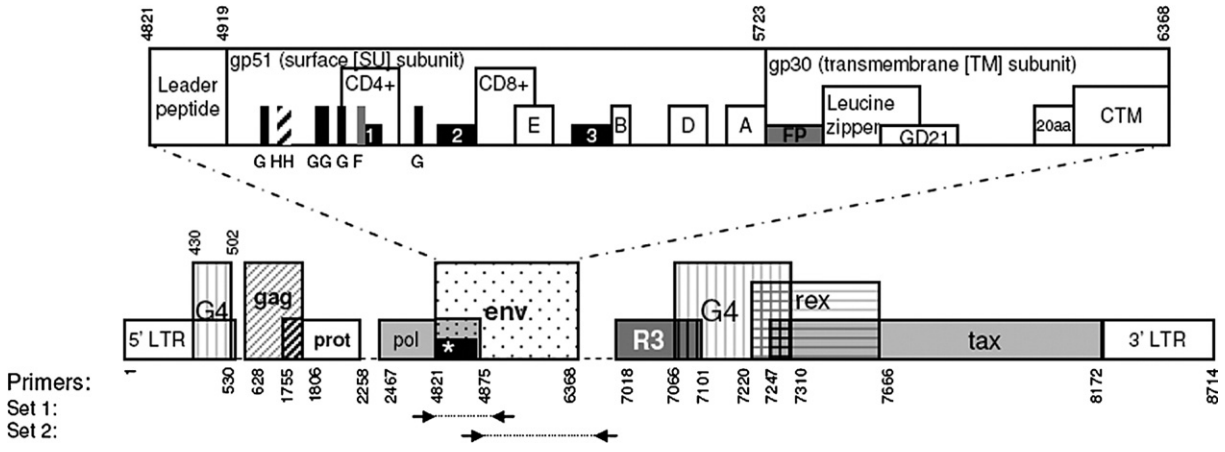


Fig. 1. The BLV genome. The length of each genomic region is not drawn to scale. The labeled rectangles in the upper panel (the *env* gene) refer to the coding sequences of corresponding Env epitopes/functional domains. The solid black rectangles labeled 1, 2, and 3 refer to the coding sequences of neutralization domains 1, 2, and 3. See Fig. 2 and Table 2 for more details of the amino acid sequences and locations of Env epitopes/functional domains. In the lower panel, the solid black rectangle labeled with an asterisk refers to the coding region shared by *env* with the 5′ *R3* and *rex* genes (nt4821–4871), and the 5′ *tax* gene (nt4868–4871). The locations of PCR-amplified *env* segments are indicated at the bottom. Annotations: 20aa: coding sequence of the uncharged 20 amino acids; A, B, D, E: linear epitopes; $CD4^+$: $CD4^+$ T cell epitope; $CD8^+$: $CD8^+$ T cell epitope; CTM: cytoplasmic domain of the transmembrane protein; *env*: *env* gene encoding viral envelope protein subunits; F, G, H: conformational epitopes; FP: fusion peptide; *gag*: *gag* gene encoding viral nucleocapsid protein subunits; LTR: long terminal repeats; *pol*: retroviral polymerase gene encoding reverse transcriptase; *prot*: *prot* gene encoding protease protein; *tax*, *rex*, *G4*, and *R3*: genes encoding BLV regulatory proteins, all having two (5′ and 3′) coding regions as shown.

sequencing, of partially overlapped genomic segments (Fig. 1; Table 1). In addition, we included in our analyses 16 full-length BLV *env* sequences already reported by other research groups and available in the GenBank database. Out of those 44 full-length BLV *env* sequences, 271 of 1548 nucleotide sites (17.51%) showed a point variation from the consensus in at least one isolate. The nucleotide variability (number of variations per site) in an individual isolate ranged from 0.064% (1/1548; CR AG-2) to 4.005% (62/1548; CR LV), with an average at 1.66% (25.68/1548). For the 28 newly sequenced isolates, their *env* gene showed nucleotide variability of 0.064% to 4.005%, with an average at 1.49%.

Forty-four deduced full-length amino acid sequences of BLV Env were aligned with their consensus sequence, with annotation of the functional domains identified to date (Fig. 2). Some large tracts of amino acid sequences, e.g. $CD4^+$ and $CD8^+$ T cell epitopes in SU and 4-3 heptad repeat and GD21 in TM, were highly conserved among isolates from different countries (Fig. 2) and were all located to the outward surface of the corresponding Env subunits in our conformational models (data not shown), thus might be of importance in designing *env* peptide-derived BLV vaccines (Daniel et al., 1993; Hadlock et al., 1995). Some small Env amino acid variations were shown to induce incomplete cross-reactivity among HTLV-1 strains and induce specific neutralizing antibodies (Blanchard et al., 1999), suggesting the important application of elucidating deltaretroviral sequence variability, viz. vaccine development and disease control. In our case, the BLV Env amino acid sequences were deduced from corresponding nucleotide sequences determined in seropositive cattle. Further information, e.g. whether sera from those animals are cross-reactive, are capable of recognizing heterogeneous Env regions, or are targeted to the tracts of conserved amino acid sequences, would be helpful in determining the possible applications of those BLV Env variability data.

In this study, variations in the deduced amino acid sequence were observed within, as well as outside, BLV Env functional domains. Some of the amino acid variations showed a trend of geographic clustering, e.g. I15V in only and most Costa Rican isolates (Fig. 2). To better evaluate the clustering pattern of BLV Env/*env* variations, we further did phylogenetic analyses and genotyping.

The phylogenetic clustering, genotyping, and geographic distribution of BLV env variations matched each other in most cases

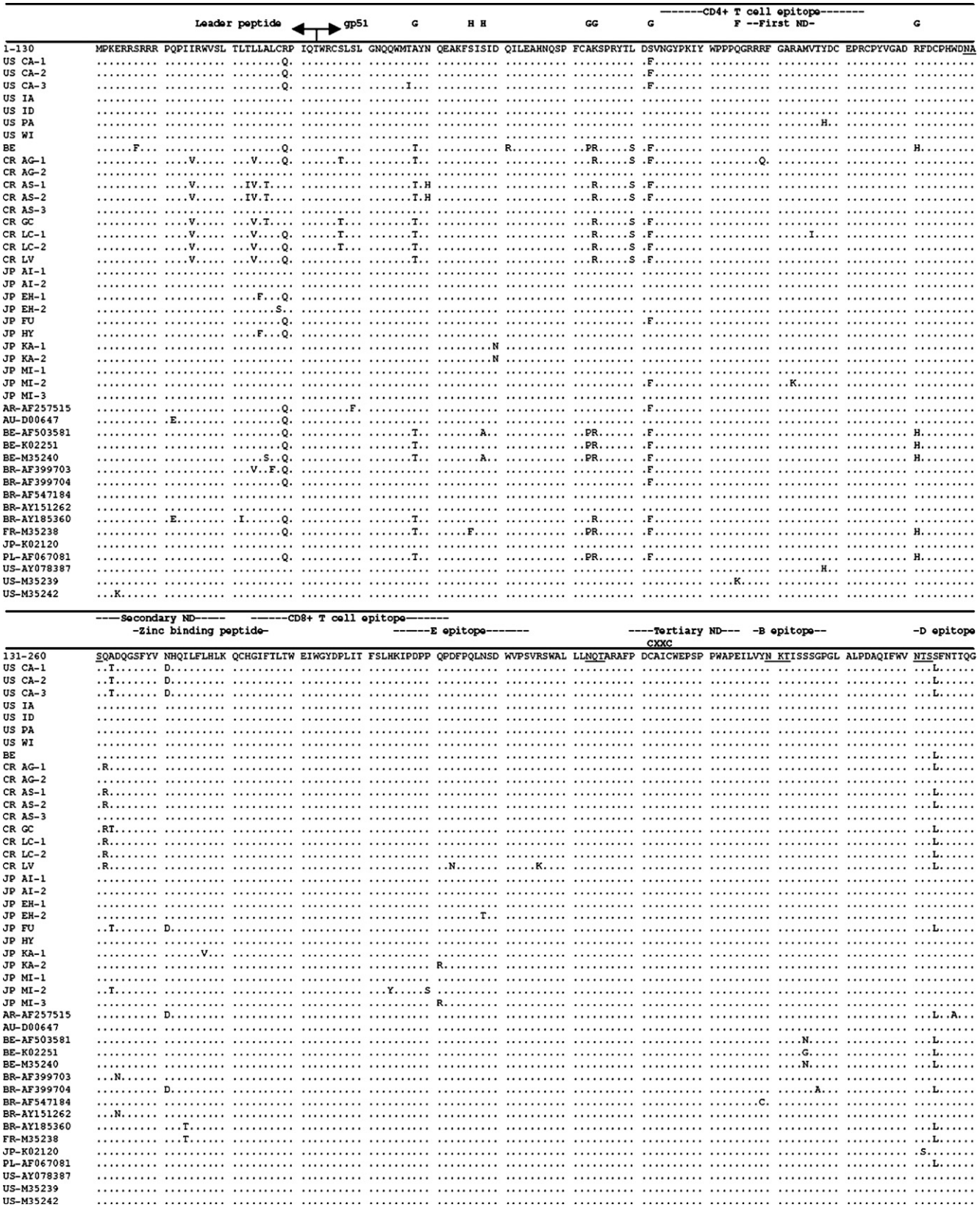
The phylogenetic tree based on full-length BLV *env* nucleotide sequences (neighbor-joining method with 1000 bootstrapping replications, MEGA 3.1) showed the grouping of isolates into 4 main clusters located in different areas worldwide: US Californian, European, Costa Rican, and consensus (showing least variation from consensus sequence) (Fig. 3). The phylogenetic tree based on deduced full-length BLV Env amino acid sequences with neighbor-joining method showed the same clustering and very similar bootstrapping values (data not shown). Trees obtained by the Minimum-Evolution method in MEGA 3.1 showed exactly the same topology (data not shown). The phylogenetic clustering roughly matched the geographic distribution of the donor cattle of BLV isolates (viral sources) (Fig. 3), consistent with the idea of BLV transmission via close contact among cattle, and reflective of the general migration trend of BLV hosts among different areas. For example, Japan used to import cattle from the USA, some of which could have been already BLV-positive at the time of importation. Consistently, the US and Japanese BLV isolates were phylogenetically grouped together, in either the US CA cluster or the consensus cluster. Seven among the nine Costa Rican isolates have consistently constituted a separate monophyletic group, based on BLV *env*/Env in this study, LTR (Zhao et al., 2007a), or regulatory gene/protein sequences (Zhao et al., 2007b). We currently do not have any idea about the reason why the other two Costa Rican isolates (CR AG-2 and CR AS-3) were consistently excluded from the CR monophyletic cluster, other than proposing that their host cattle might be recent additions to the herds close to the time of sample collection, and the BLV genotypes were more consistent with the original parental herds than the new herds to which the animals were transferred.

Genotyping based on restriction fragment length polymorphism (RFLP) has been frequently used for a general idea on genomic composition when detailed sequence information is not available. Genotypes using deduced RFLP on 444-bp partial BLV *env* sequence (nt5099–5542), by 5 restrictive enzymes as described previously (Fechner et al., 1997), were determined for each isolate (Table 1). In another previous study (Licursi et al., 2002), although RFLP was performed using the same five restriction enzymes, only three (*Bcl*I, *Hae*III, and *Pvu*II) were used for genotyping. RFLP, although based on partial sequence in these cases, can provide supplementary information to full length-based phylogenetic clustering. For example, isolate JP-K02120 (Sagata et al., 1985) was determined to be genotype B/3 together with JP MI-1, -2 and -3 (isolates from this study), unlike the other isolates in the consensus group (genotype C/1) (Table 1). Coincidentally, those four isolates were collected from geographically adjacent prefectures in Japan (Iwate and Miyagi). To evaluate how well the two clustering methods, the phylogeny based on the full-length BLV *env* gene vs. genotyping based on partial *env* sequences (Fechner et al., 1997), matched each other, we further calculated the percentage of "matched pairs". For each random pair of isolates, if they were clustered together by both methods or if they were consistently separated into two groups by those two methods, they were defined as a "match pair". Out of all 946 ($=44\times43/2$) possible pairs of the 44 isolates, 747 pairs (78.96%) matched, supporting the idea that those two clustering methods (phylogeny and genotyping) corresponded in most cases (Fig. 3; Table 1).

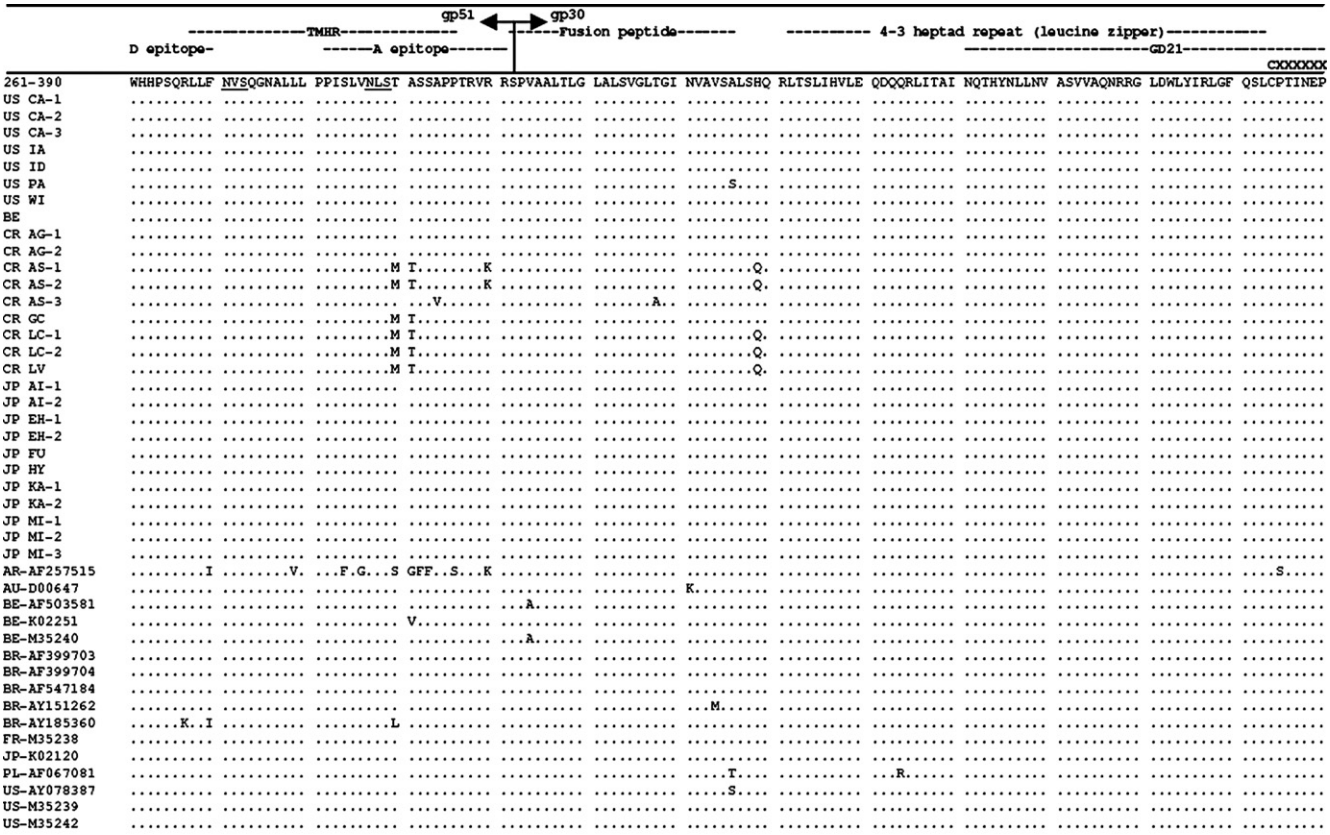
Selection pressures were suggested for BLV env, both globally and locally

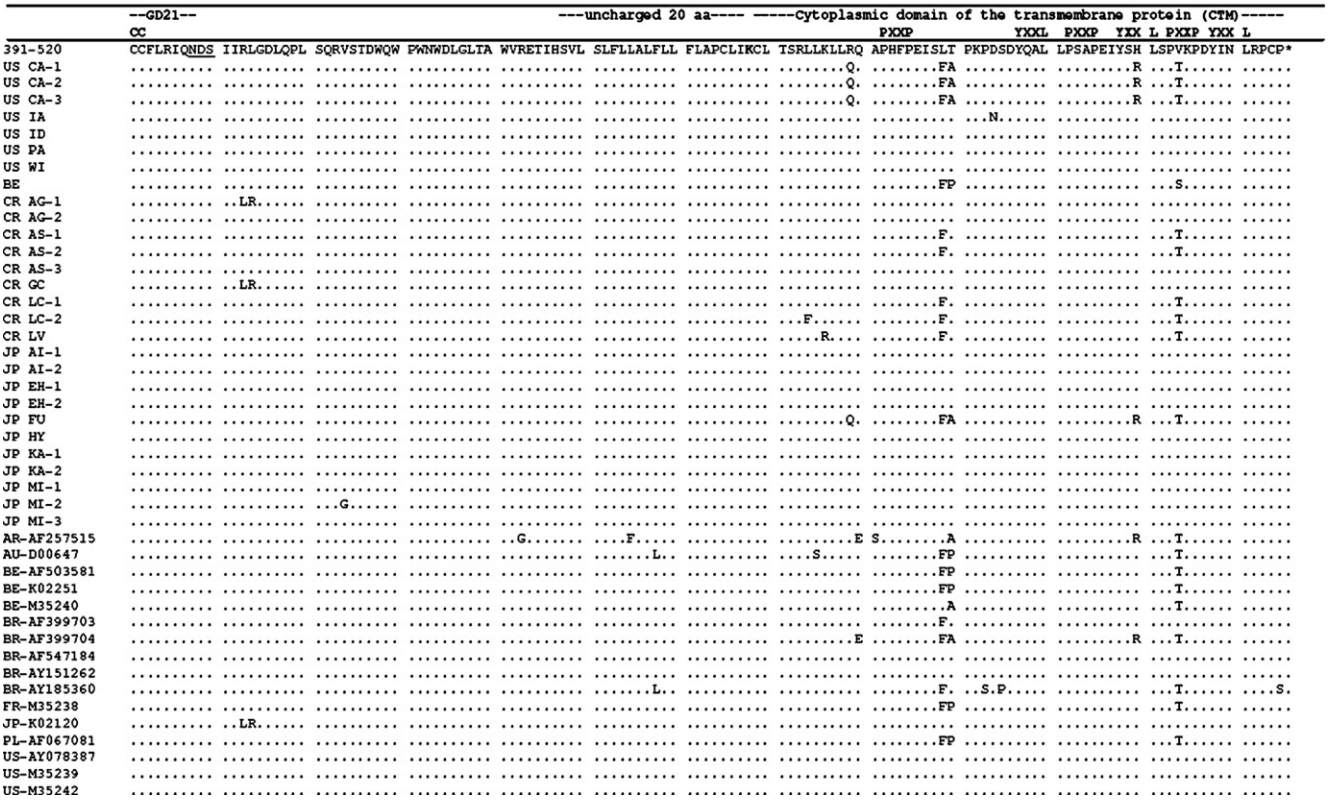
For the 44 full-length BLV *env* genes, the difference in the overall nucleotide variability (number of variations per nucleotide per isolate) among the leader peptide, gp51, and gp30 sub-

coding regions was moderately significant (Friedman's test with Dunn's post test, $p=0.0266$). We further conducted selection tests on the full-length *env* and on each individual sub-region. When the rate of non-synonymous substitution (P_N = number of non-synonymous substitutions per non-synonymous site) versus the rate of synonymous substitution (P_S = number of



synonymous substitutions per synonymous site) were compared (Wilcoxon matched pairs test), significant differences were detected in total *env* ($p<0.0001$) as well as in leader peptide ($p<0.005$), gp51 ($p<0.0001$), and gp30 ($p<0.0001$) sequences. The ratios of P_N/P_S were calculated for evidence of selection (<1 for negative and >1 for positive selection). Negative





selection selectively removes rare mutations that are deleterious. It can result in the maintenance of conserved genomic sequences over long periods of evolutionary time. In comparison, positive selection increases the frequency of certain variations. The BLV *env* gene as a whole and all its sub-coding regions (leader peptide, gp51, and gp30) showed negative selection ($P_N < P_S$) (Table 2).

The possible selections on the BLV Env epitopes/functional domains identified to date are summarized by Table 2. Several coding sequences for functional domains (epitope F, zinc binding cysteine, WAPE motif, CXXC motif, and CX_6CC motif) did not show a statistically significant selection (Table 2), probably due to short sequence lengths (3–12 nucleotides). Those coding sequences all showed reduced nucleotide diversity and amino acid diversity (P_N) compared with those in their sequence context (gp51 or gp30 coding region), consistent with the need of structural conservation in those functional domains.

Negative selection was observed in the coding sequences for most functional domains (Table 2), as expected for their importance in sequence and conformational conservation. For example, the gp51 zinc binding peptide, which interacts with zinc ion and is indispensable in in vivo viral fusion (Gatot et al., 2002), showed a mean P_N/P_S at 0.1946 ($P<0.005$; Table 2). Another example of stringently conserved gp51 motifs is the specific N-linked glycosylation sites. The glycosylation modifications in retroviral SU subunits can help virus escape the recognition by neutralizing antibodies, by masking principal neutralizing epitopes (Chackerian et al., 1997; Cheng-Mayer et al., 1999; Johnson et al., 2003). This idea might also be applicable in the case of BLV Env, as indicated in our conformational model (Fig. 4A), and is consistent with the dramatically reduced P_N/P_S at 0 ($P<0.001$) observed in this study (Table 2). In the BLV TM subunit gp30, a stretch of 20 uncharged residues (aa438–457), which is believed to span host cell membranes (Rice et al., 1984), showed a P_N/P_S of 0.0325 ($P<0.001$; Table 2). Furthermore, all of the predicted amino acid substitutions (L445F and F448L) within that uncharged 20aa stretch were also uncharged, sustaining its hydrophobicity. Interestingly, K458, the first positively charged residue following that stretch as a signal to stop the transfer of peptide chains through a lipid membrane, showed no variation in any sample. The same conservation is also observed for R463, the second, "back-up" positive residue downstream to that 20aa stretch. Other examples of highly conserved TM motifs include the YXXL sequences, required for BLV viral entry and incorporation of viral Env protein into virions (Inabe et al., 1999), as well as involved in the down-regulation of the surface exposure of BLV Env, assisting in viral evasion of adaptive immunity (Novakovic et al., 2004). Correspondingly, YXXL motifs showed a P_N/P_S at 0 ($P<0.005$; Table 2).

Positive selection was observed for immunostimulatory epitopes D/D' and G (Table 2). The extremely high P_N/P_S value of G epitope is consistent with its role as a critical epitope involved in BLV infectivity, syncytium induction, and immune escape. Previous studies also suggested that gp51 sequence variability might preferentially involve structural changes underlying viral escape from immune surveillance (Molteni et al., 1996). Interestingly, some BLV Env variants that showed reduced binding by monoclonal antibodies specific for the conformational epitopes F, G, and H did not show appreciably reduced syncytium induction (Johnston et al., 2002), implying that immuno-escape mutants could have intact pathogenicity. Similar to epitope G, conformational epitope F also showed a P_N higher than P_S (Table 2), but the difference was not statistically significant, probably due to short sequence length (3 nucleotides). In comparison, the third conformational epitope H showed negative selection. Such a difference among the observed selection patterns of epitopes F, G, and H is not due to overlapped reading frames, which affect the selection patterns in the pXBL coding sequences of BLV (Zhao et al., 2007b). Our survey on the charges and hydropathy index values of the amino acids constituting or flanking those conformational epitopes did not reveal any special trend, either. Currently we do not have any explanation for that difference, other than proposing that such observed negative selection on H conformational epitope might be artifactual due to its short length (6 nucleotides).

The cluster-specific consensus sequences differed from the consensus calculated for the entire sample pool (Fig. 2), thus we further examined the possible selection in BLV *env* for each individual phylogenetic group (Fig. 3). Most cluster-specific analyses showed negative selection (Table 3), the same as for the entire population. The comparison of relative variability among leader peptide, gp51, and gp30, in the forms of either overall nucleotide diversity or amino acid diversity (P_N), showed variable results in different phylogenetic clusters (Table 3), depending on the isolates included for analyses. This may help explain the previous concern of the apparently controversial results described by different research groups (Camargos et al., 2006), where higher variability were observed for leader peptide region in some studies but for gp51 region in others.

The BLV gp51 and gp30 conformational models map most natural variations on the surface and spatially close to functional domains

Advances in research on retroviral pathogenesis and vaccine development require a better understanding of the antigenic structure of envelope surface protein which is directly involved in infectivity events, such as cell surface receptor recognition and binding. To understand how the Env

Fig. 2. The deduced Env amino acid sequences of 44 BLV isolates. In each panel, the 28 full-length BLV Env sequences predicted by the nucleotide sequences determined in this research are shown above the 16 already published BLV Env sequences. The consensus sequence is shown at the top of each panel. Dots represent where the amino acids match the consensus sequence. Amino acid 1 indicates the start of the BLV Env. Labeled horizontal lines indicate the position of identified epitopes and functional domains. Some epitopes (F, G, and H) consist of a single amino acid. AR: Argentina; AU: Australia; BE: Belgium; BR: Brazil; CR: Costa Rica; F, G, H: conformational epitopes; FR: France; JP: Japan; ND: neutralization domains; PL: Poland; TMHR: transmembrane hydrophobic region; US: the United States.

Table 1
BLV isolates

Samples	Year	Viral source	PC	Genotype	
				F	L
US CA-1	**1997***	Davis, California, USA	US CA	G′	5
US CA-2	**1997***	Davis, California, USA	US CA	G′	5
US CA-3	**1997***	Davis, California, USA	US CA	G′	5
JP FU	**2004**	Fukuoka, Japan	US CA	G′	5
JP AI-1	**2004**	Aichi, Japan	CONS	C	1
JP AI-2	**2004**	Aichi, Japan	CONS	C	1
JP EH-2	**2004**	Ehime, Japan	CONS	C	1
JP MI-1	**2004**	Miyagi, Japan	CONS	B	3
JP MI-2	**2004**	Miyagi, Japan	CONS	N/A	3
JP MI-3	**2004**	Miyagi, Japan	CONS	B	3
US IA	**1975***	Iowa, USA	CONS	C	1
US ID	**1994***	Idaho, USA	CONS	C	1
US WI	**1991***	Wisconsin, USA	CONS	C	1
US PA	**1976***	Pennsylvania, USA	CONS	C	1
US-AY078387	2002	USA	CONS	C	1
US-M35239	1990	USA	CONS	C	1
US-M35242	1990	USA	CONS	C	1
JP KA-1	**2004**	Kamikawa, Japan	CONS	C	1
JP KA-2	**2004**	Kamikawa, Japan	CONS	C	1
JP-K02120	1985	Iwate, Japan	CONS	B	3
BR-AF547184	2002	Minas Gerais, Brazil	CONS	C	1
BR-AY151262	2002	Minas Gerais, Brazil	CONS	C	1
CR AG-2	**2003***	Guanacaste-Canas, CR	CONS	C	1
CR AS-3	**2003***	Alajuela-San Carlos, CR	CONS	C	1
JP EH-1	**2004**	Ehime, Japan	CONS	C	1
JP HY	**2004**	Hyogo, Japan	CONS	C	1
FR-M35238	1990	France	EU	A	6
PL-AF067081	1998	Poland	EU	A	6
BE-K02251	1984	Belgium	EU	A	6
BE-AF503581	1993	Belgium	EU	A	6
BE-M35240	1990	Belgium	EU	A	6
BE	**2003**	Belgium	EU	A	6
CR AG-1	**2003***	Alajuela-Grecia, CR	CR	A	6
CR GC	**2003***	Guanacaste-Canas, CR	CR	N/A	N/A
CR AS-1	**2003***	Alajuela-San Carlos, CR	CR	A	6
CR AS-2	**2003***	Alajuela-San Carlos, CR	CR	A	6
CR LV	**2003***	Limon-Valle de la Estrella, CR	CR	A	6
CR LC-1	**2003***	Limon-Central, CR	CR	A	6
CR LC-2	**2003***	Limon-Central, CR	CR	A	6
BR-AF399703	2001	Northern Paran, Brazil	MISC	C	1
AU-D00647	1990	Australia	MISC	C	1
BR-AY185360	2002	Mato Grosso do Sul, Brazil	MISC	C	1
AR-AF257515	2000	Argentina	MISC	A	6
BR-AF399704	2001	Minas Gerais, Brazil	MISC	A	6

The full-length env sequences from the 28 bolded isolates were determined in this research, while the other 16 are all the full-length BLV env published by other researchers. The geographic locations (viral source) of donor cattle and phylogenetic clustering (PC) of each isolate are annotated. The years when the proviral BLV-containing genomic DNA samples were collected directly from donor cattle (bolded and asterisked), obtained as gifts from other researchers (bolded), or when the proviral sequences were published in GenBank (regular font) are indicated. Isolates US IA and US PA were obtained from cell lines established from BLV-positive cattle and maintained in our laboratory, and cellular DNA samples containing BLV provirus were collected in 2005 and 1996, respectively. Predicted restriction fragment length polymorphism (RFLP) used the five restrictive enzymes described before (Fechner et al., 1997). Genotypes were determined using RFLP from all five enzymes (Fechner et al., 1997; column "F") or from *Bcl*I, *Hae*III, and *Pvu*II (Licursi et al., 2003; column "L") as described previously.
AR: Argentina; AU: Australia; BE: Belgium; BR: Brazil; CONS: consensus cluster; CR: Costa Rica; EU: European cluster; F: genotypes according to Fechner et al., 1997; FR: France; JP: Japan; L: genotypes according to Licursi et al., 2002; MISC: miscellaneous samples; N/A: not applicable; PC: phylogenetic clustering; PL: Poland; US: the United States; US CA: US Californian cluster.
Samples with an accession number as part of the name are already published sequences, and their geographic origins were indicated by two-letter abbreviations for countries.

residue substitutions might affect its configuration, we first made a model of the BLV gp51 N-terminus (aa34–186) which holds the putative receptor binding domain (RBD), by following the homology alignment strategy developed by Johnston et al. (2002) but using different software for threading, viewing, and annotation. The proposed conformation (Fig. 4A) was consistent with that predicted previously (Johnston et al., 2002). All the biochemically identified gp51

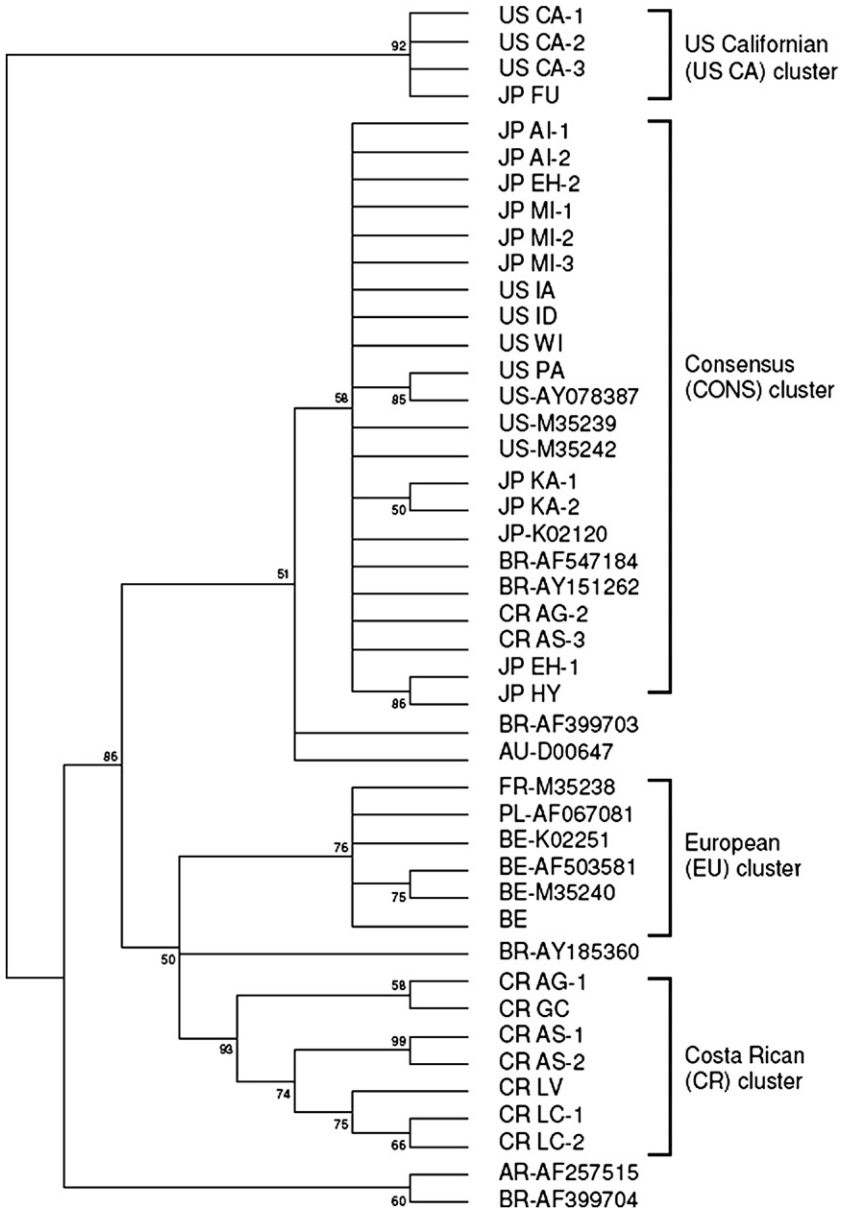


Fig. 3. Clustering of BLV isolates based on 44 full-length BLV *env* sequences. The phylogenetic tree shown here was constructed with the Neighbor-Joining method in MEGA 3.1, checked with 1000 bootstrapping replications, and condensed with cut-off value at 50%. The numbers on forks indicate the percentage of times the group consisting of the species to the right of that fork occurred among the trees, out of 1000 trees. See Table 1 for geographical origin of each isolate.

N-terminal epitopes (E, F, G, H, $CD4^+$ T cell, $CD8^+$ T cell) and functional domains (neutralization domain [ND] 1, ND2, zinc binding peptide) were mapped to the surface (Fig. 4A and data not shown), and some (e.g. $CD4^+$ and $CD8^+$ T cell epitopes) constituted joining reactive surface regions, despite their lack of continuity in primary structure (amino acid sequence) (data not shown). The calculated potential N-glycosylation site (aa129–131) was also mapped to surface, and its proposed additional carbohydrate could mask the ND2 (Fig. 4A). It is consistent with the predicted glycosylation-mediated neutralization escape in simian immunodeficiency virus (SIV) via shielding of principal neutralization sites (Cheng-Mayer et al., 1999). SIV mutants with elimination of N-glycosylation sites in their gp120 SU showed dramatically increased sensitivity to neutralizing antibodies, partially by exposing localized epitopes (Johnson et al., 2003). In comparison, another candidate BLV N-glycosylation site aa67–69, which was rejected by NetNGlyc calculation, was mapped to the face opposite of the neutralization domains (data not shown), showing consistence between those two computerized predictions.

In this model, all the three conformational epitopes (F, G, H), both ND1 and ND2, and N-glycosylation site are primarily mapped to one face of the molecule (determined as front face) or to the edge of the opposite face (back face). Those occupied surface areas, as proposed by Johnston et al. (2002), may be involved in interactions with cellular surface receptors and neutralization antibodies, thus involved in cell entry (infection) and immunogenicity. All the 17 polymorphic informative (variables occurring in a population over one time) N-terminal

Table 2
Tests for evidence of selection on BLV env sequences encoding epitopes and functional domains

Peptides	Domains		Location (aa #)	References	Nucleotide variability	P_N	P_S	P_N/P_S	p values ($P_N \neq P_S$)
Full-length	Overall		1–515		0.0173	0.0068	0.0489	0.1391	<0.001
Leader Peptide	Overall		1–33		0.0199	0.0165	0.0289	0.5709	<0.005
gp51	Overall		34–301		0.0186	0.0067	0.0564	0.1188	<0.001
	Immunostimulatory	A	282–301	1, 3	0.0192	0.0155	0.0270	0.5741	<0.005
	epitopes	B, B′	228–238	1, 3	0.0088	0.0048	0.0199	0.2412	<0.05
		D, D′	251–270	1, 3	0.0121	0.0128	0.0095	1.3474	<0.05 *
		E	175–194	1, 3	0.0444	0.0024	0.2040	0.0118	<0.001
		F	95	3, 12, 15	0.0105	0.0118	0.0000	N/A	NS
		G	48, 73, 74, 82, 121	3, 12, 15	0.1339	0.1736	0.0314	5.5287	<0.001*
		H	56, 58	3, 12, 15	0.0570	0.0201	0.1311	0.1533	<0.02
	Neutralization domains	1	97–106	2, 5	0.0150	0.0033	0.0439	0.0752	<0.001
		2	131–150	3, 2, 5	0.0182	0.0119	0.0409	0.2910	<0.001
		3	210–225	3, 2, 5	0.0069	0.0000	0.0310	0.0000	<0.005
	T cell epitopes	$CD4^+$	84–113	5	0.0134	0.0021	0.0446	0.0471	<0.001
		$CD8^+$	154–183	4, 5	0.0181	0.0017	0.0720	0.0236	<0.001
	Zinc-binding peptide		137–156	8	0.0079	0.0043	0.0221	0.1946	<0.005
	Zinc-binding cysteine		152	8	0.0076	0.0000	0.0455	0.0000	NS
	GYDP strong turn		164–167	6	0.0048	0.0000	0.0174	0.0000	NS
	TMHR		268–297	5	0.0225	0.0096	0.5631	0.0170	<0.001
	CXXC in disulfide bond		212–215	10	0.0000	0.0000	0.0000	N/A	NS
	WAPE tetrapeptide		221–225	3	0.0056	0.0000	0.0289	0.0000	NS
	SSSG strong turn		234–237	6	0.0083	0.0115	0.0000	N/A	NS
	Potential glycosylation sites		129–131, 203–205, 230–232, 251–253, 271–273, 287–289, 398–400	13	0.0332	0.0000	0.1493	0.0000	<0.001
gp30	Overall		302–515		0.0156	0.0056	0.0442	0.1267	<0.001
	Fusion peptide		302–326	11	0.0163	0.0036	0.0433	0.0831	<0.001
	4-3 Heptad repeat		332–383	7	0.0106	0.0002	0.0407	0.0049	<0.001
	GD21		351–398	7	0.0169	0.0002	0.6833	0.0003	<0.001
	CX_6CC in disulfide bond		384–392	10	0.0062	0.0000	0.0375	0.0000	NS
	Uncharged 20aa tract		438–457	16	0.0155	0.0016	0.0493	0.0325	<0.001
	CTM		459–515	14	0.0232	0.0167	0.0411	0.4063	<0.001
	PXXP in CTM		472–475, 492–495, 503–506	14, 17	0.0228	0.0000	0.0685	0.0000	<0.001
	YXXL signaling motifs		487–490, 498–501, 508–511	9, 14, 17	0.0166	0.0000	0.5260	0.0000	<0.005

The p (probability) values ($T \leq t$) were determined using Wilcoxon matched pairs test.
NS: not significant ($p > 0.05$).
Reference numbers represent: 1: Callebaut et al. (1991); 2: Callebaut et al. (1993); 3: Camargos et al. (2006); 4: Daniel et al. (1993); 5: Dube et al. (2000); 6: Gallaher et al. (1995); 7: Gatot et al. (1998); 8: Gatot et al. (2002); 9: Inabe et al. (1999); 10: Johnston et al. (2002); 11: Kobe et al. (1999); 12: Mamoun et al. (1990); 13: NetNGlyc prediction web server 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/); 14: Novakovic et al. (2004); 15. Portetelle et al. (1989); 16: Rice et al. (1984); 17: Willems et al. (1995).
* p values where P_N (rate of non-synonymous substitution) was significantly higher than the rate of synonymous substitution P_S (positive selection); otherwise, P_N was significantly lower than P_S (negative selection).

variations (aa37, 48, 50, 58, 60, 73, 74, 80, 82, 108, 121, 132, 133, 134, 141, 144, and 181; Fig. 2) were mapped to the surface, either on the front face or at the edge of the back face, and most (except aa48) were overlapping with or spatially adjacent to certain gp51 N-terminal immunostimulatory epitopes, indicating possible biological importance of those frequent natural variations, e.g. ensuing conformational variability for immune escape. A parallel model built with substitutions of the most frequent natural variations (A48T, A73P, K74R, L80S, S82F, R121H, Q132R, A133T, and N141D; Fig. 2) revealed subtle conformational difference from the wild type (Figs. 4A, B). Further investigations would be needed to determine the possible biological significance of such conformational difference.

An extensive functional study on BLV gp51 N-terminus using introduced single-amino acid mutations demonstrated that all of the functionally detrimental mutations mapped to the front face (Johnston et al., 2002), with most of them close to or within a gp51 functional domain, further supporting the biological significance of the residues on the front face. However, none of those detrimental, experimentally introduced mutations occurred naturally in our sample pool which includes all known full-length BLV *env* sequences to date, neither were they located adjacent to the frequent natural variations based on our structural model (data not shown). It is consistent with the idea that function-abolishing mutations may be selected against in the natural environment.

We also showed a model for the BLV gp30 external domain (aa327–415) using homology alignment between HTLV-1 gp21 and BLV gp30 described by Kobe et al. (1999) and homology modeling based on the crystal structure of HTLV-1 gp21

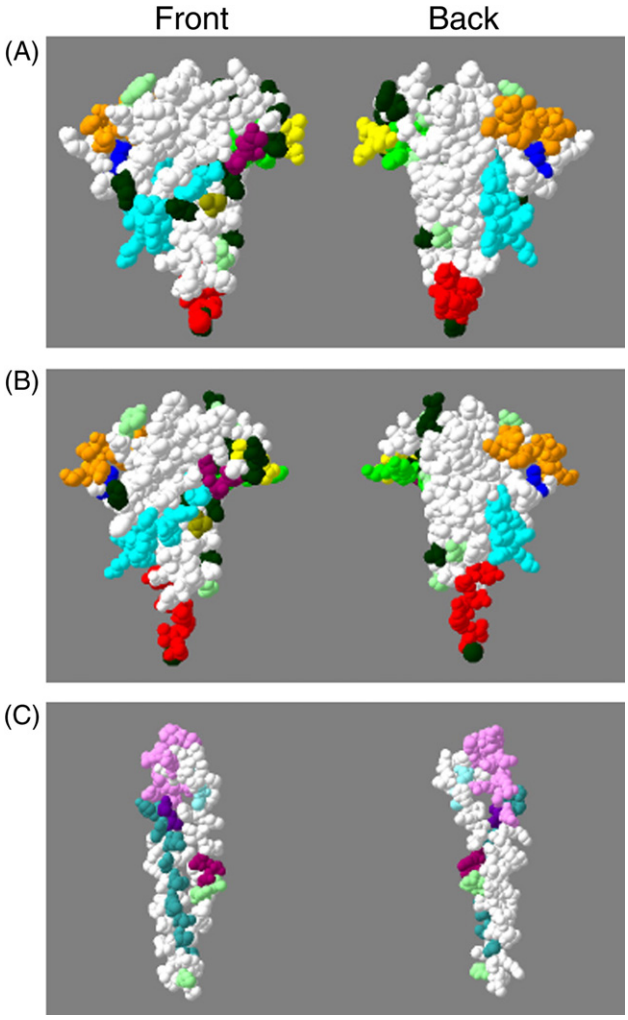


Fig. 4. Proposed space-filling models for the BLV gp51 N-terminus (A, B) and gp30 external domain (C). Left column=front views, right column=back views. (A) Model of the BLV gp51 N-terminus (aa34–185), showing locations of gp51 N-terminal epitopes, functional domains, and polymorphic informative (PI) variation sites. Color annotations: E epitope: red; F epitope: blue; glycosylation site: magenta; H epitope: olive green; ND1: orange; ND2: yellow; PI variation sites: light green if not overlapped with epitopes, dark green if overlapped; zinc binding peptide: turquoise. All G epitopes are among the PI variation sites and shown in dark green. Bright green refers to the overlapped region between ND2 and zinc binding peptide. (B) Model of the BLV gp51 N-terminus (aa34–185) with residue substitutions of the most frequent natural variations (A48T, A73P, K74R, L80S, S82F, R121H, Q132R, A133T, and N141D). Locations of the same epitopes, functional domains, and polymorphic informative (PI) variation sites are shown as in (A). (C) Model of the BLV gp30 external domain (aa327–415). Color annotations: 4-3 heptad repeat (leucine zipper): blue green; experimentally introduced fusion-defective, infectivity-intact mutations (Gatot et al., 1998): purple; glycosylation site: magenta; immunodominant cysteine residues: light blue; immunosuppressive region: pink; PI variation sites: light green (none overlapped with epitopes).

ectodomain crystallized as a maltose-binding protein chimera (Kobe et al., 1999). The conformation we propose (Fig. 4C) corresponds to the BLV TM model predicted using the hydrophobic cluster analysis based on the coiled coil core described for influenza A virus hemagglutinin HA2 (Gatot et al., 1998). This region (aa327–415) could be involved in cell fusion, and the integrity of this coiled coil structure is essential for BLV Env protein multimerization (Gatot et al., 1998). Two experimentally introduced mutations (A361V and A365S; Gatot et al., 1998), which affected fusion but not infectivity in vivo, were not observed naturally in our sample pool and were mapped adjacent to the leucine zipper (Fig. 2) on our model (Fig. 4C). The predicted locations of immunosuppressive region (aa361–381), immunodominant cysteine residues (C384, C391, C392), and the polymorphic informative variable sites (aa329, 403, 404) were shown (Fig. 4C).

Natural deletion mutations were bounded by similar flanking sequences and would result in truncations and aberrant coding frames

Massive deletions in BLV *env* gene were detected in 3 Japanese isolates, where full-length *env* genes coexisted in the same sample (Fig. 5A). Those three full-length *env* sequences differed from each other and from all other *env* sequences in this study (Fig. 2), making it unlikely that they represent cross-contamination. Those natural deletions are summarized by Table 4. In isolates JP AI-1 and JP EH-2, deletions resulted in prematurely terminated gp51 transcript and questionable expression of gp30 subunit due to frame shift, despite their conserved gp30 coding sequences (Table 4; Figs. 5B, D). It might be possible that they use alternative CTG start codons (Table 4), which are readily used in BLV to initiate G4 gene transcription, to generate full-length gp30 transcripts which could then undergo cleavage at the normal cleavage site (RVRR, aa298–301 in wild type) to release wild type gp30 peptides. In another isolate JP EH-1, deletions did not generate a frame shift, thus resulting in an aberrant C-terminus of leader peptide, missed middle segment of gp51, but intact gp30 subunit (Table 4; Fig. 5C). In each of those 3 isolates, the nucleotide sequences flanking the upstream and downstream deletion sites showed a high level of sequence similarity (Table 4). Likewise, in de novo deleted feline leukemia virus (FeLV)

Table 3
Tests for evidence of selection on BLV env in each phylogenetic cluster

Cluster	ONV	P_N	P_S	P_N/P_S	p values
(A) Leader peptide					
US CA	0.0000	0.0000	0.0000	N/A	NS
CR	0.0248	0.0160	0.0489	0.3271	0.0391
CONS	0.0037	0.0038	0.0033	1.1388	NS
EU	0.0241	0.0046	0.0802	0.0576	0.0156
(B) gp51 (surface glycoprotein)					
US CA	0.0034	0.0004	0.0130	0.0313	NS
CR	0.0068	0.0042	0.0148	0.2842	0.0078
CONS	0.0029	0.0013	0.0078	0.1713	<0.0001
EU	0.0052	0.0027	0.0130	0.2089	0.0469
(C) gp30 (transmembrane glycoprotein)					
US CA	0.0039	0.0000	0.0150	0.0000	NS
CR	0.0120	0.0042	0.0343	0.1229	0.0156
CONS	0.0030	0.0008	0.0095	0.0802	<0.0001
EU	0.0140	0.0067	0.0350	0.1901	0.0156

Phylogenetic clusters are shown in Fig. 3. The p (probability) values ($T \leq t$) were determined using Wilcoxon matched pairs test.
NS: not significant ($p>0.05$); ONV: overall nucleotide variability; P_N=the rate of non-synonymous substitution; P_S=the rate of synonymous substitution.

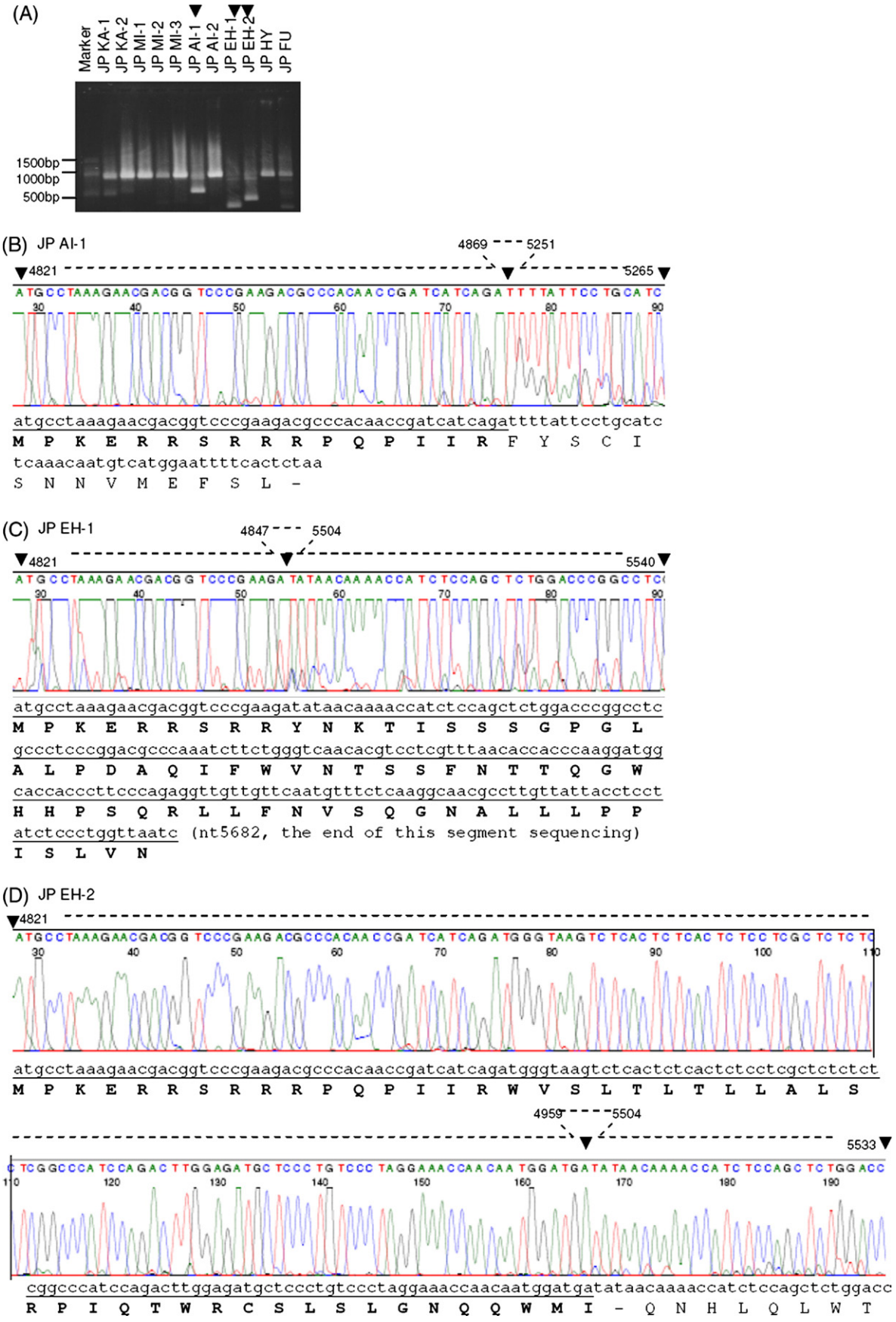

(A)
Marker
JP KA-1
JP KA-2
JP MI-1
JP MI-2
JP MI-3
JP AI-1
JP AI-2
JP EH-1
JP EH-2
JP HY
JP FU
1500bp
1000bp
500bp
(B) JP AI-1
4821
4869
5251
5265
atgcctaaagaacgacggtcccgaagacgcccacaaccgatcatcagattttattcctgcatc
M P K E R R S R R R P Q P I I R F Y S C I
tcaaacaatgtcatggaattttcactctaa
S N N V M E F S L -
(C) JP EH-1
4821
4847
5504
5540
atgcctaaagaacgacggtcccgaagatataacaaaaccatctccagctctggacccggcctc
M P K E R R S R R Y N K T I S S S G P G L
gccctcccggacgcccaaatcttctgggtcaacacgtcctcgtttaacaccacccaaggatgg
A L P D A Q I F W V N T S S F N T T Q G W
caccacccttcccagaggttgttgttcaatgtttctcaaggcaacgccttgttattacctcct
H H P S Q R L L F N V S Q G N A L L L P P
atctccctggttaatc (nt5682, the end of this segment sequencing)
I S L V N
(D) JP EH-2
4821
atgcctaaagaacgacggtcccgaagacgcccacaaccgatcatcagatgggtaagtctcactctcactctcctcgctctctct
M P K E R R S R R R P Q P I I R W V S L T L T L L A L S
4959
5504
5533
cggcccatccagacttggagatgctccctgtccctaggaaaccaacaatggatgatataacaaaaccatctccagctctggacc
R P I Q T W R C S L S L G N Q Q W M I - Q N H L Q L W T

Table 4
Natural deletions in BLV *env* gene

Sample	JP AI-1	JP EH-1	JP EH-2
Donor cattle breed	Holstein	Japanese black	Japanese black
WBC at sample collection	9200 cells/mm^3	14,100 cells/mm^3	13,500 cells/mm^3
Deleted region	nt4870–5251	nt4848–5504	nt4960–5504
Deleted length	382nt	657nt	545nt
Upstream flanking sequence (UFS)	nt4860–4869: atcatcagat	nt4837–4847: ggtcccgaaga	nt4959–4970: acagcatataac
Downstream flanking sequence (DFS)	nt5242–5251: atcatcagat	nt5486–5504: ggctcccgaaatattagta	nt5501–5510: agtatataac
UFS/DFS identity	10/10	9/10	9/10
UFS/DFS unmatched gap	0	8	3
Leader peptide (LP)	C-terminus missing	C-terminus missing	Intact
gp51	N-terminus missing, FS, and PMT	N-terminus missing, C-terminus intact	Middle part missing, FS, and PMT
gp30	CNS intact	Intact	CNS intact
In-frame ATG for gp30	None	nt4821–4823	None
1st in-frame CTG for gp30	nt5322–5324	N/A	nt5673–5675
pol stop codon	Missing; (nt5315–5317)	Missing; (nt5740–5742)	Intact
pol encoded product	Extra 20 amino acids	Extra 70 amino acids	Intact
rex/R3 5′ coding region	Missing 3′ 2nt	Missing 3′ 24nt	Intact
rex/R3 SDS	Missing; (nt5275–5277)	Missing; (nt5568–5570)	Intact
tax 5′ coding region	Missing start codon	Missing entire region	Intact
tax SDS	N/A	N/A	Intact

Three samples with a majority of SU (surface subunit) PCR products shorter than consensus are summarized. They also all showed coexisting full-length *env* sequences (Fig. 5A).
CNS: coding nucleotide sequence; FS: frame shift; PMT: premature termination; SDS: splicing donor site; WBC: white blood cells.
First in-frame CTG alternative start codon refers to the first in-fame CTG triplet after the deletion site. Nucleotide positions in parentheses refer to the first possible downstream alternative sites if the original sites are missing.

env sequence isolated from FeLV-induced tumor, both ends of the deletion were flanked by the same 6nt sequence (Shi and Roy-Burman, 2000). This probably underlies a process of internal deletion based on recombination induced by sequence similarity; however, more evidence is needed to test this possibility.

Deletions in the *env* region have been reported in endogenous retrovirus-like sequences, e.g., endogenous feline retroviral species in cats (Reeves and O'Brien, 1984; Tchenio and Heidmann, 1991). These have also been detected in other retroviruses and were related to functional defectiveness, e.g. an HIV-1 Env mutant with a deletion in its membrane-proximal domain did not mediate cell fusion (Dimitrov et al., 2003). Deletions of partial genome, especially *env* gene, have been more frequently reported in leukemia/lymphoma-inducing viruses, and such natural defective viral forms have been related to malignant transformation. Examples include Rous sarcoma virus with completely deleted *env* detected in chicken sarcoma (Nunn et al., 1984), myoloproliferative leukemia virus with deletion and rearrangement in *env* region detected in acute leukemia (Penciolelli et al., 1987), Friend spleen focus-forming virus with partial deletion and frame-shifting insertion in *env* region detected in acute erythroleukemia in mice (Watanabe et al., 1995), avian leucosis virus with multiple partial deletions including *env* region detected in leukemia and bursal lymphoma (Westaway et al., 1984; Goodenow and Hayward, 1987), and feline leukemia virus with *env* truncation and recombination detected in thymic and alimentary lymphosarcoma (Sheets et al., 1993; Shi and Roy-Burman, 2000). For HTLV-1, which belongs to the same genus *Deltaretrovirus* as BLV, proviral deletions in *env* region have been detected in peripheral blood cells of patients with T-cell prolymphocytic leukemia (Kojima et al., 1998), with adult T-cell leukemia/lymphoma (Ohshima et al., 1993), or with mycosis fungoides (Hall et al., 1991). Although in previous studies, deleted mutants were found coexisting with wild type viruses in tumor cells and not all the tumor cells examined contained deleted mutants, those studies still suggested a possible correlation between the *env* deletion mutations and viral oncogenesis/leukemogenesis. For BLV, natural *env* deletions have been rarely reported. One previous report described a deletion around *env* region in infected cells from enzootic bovine leucosis (Ogawa et al., 1987).

Deletion and associated frame shift could severely affect the amino acid sequence, and thus the conformational structure, of protein product. The three natural deletions observed in this study would abolish the gp51 N-terminus RBD, whose wild type conformational models are predicted in Fig. 4A. Altered Env conformation or lack of Env expression on the surface of

Fig. 5. Natural deletion mutations in N-terminal Env transcripts. (A) Representative gel of PCR amplification for the genomic segment encoding leader peptide and gp51 (set 1 primer illustrated in Fig. 1). Expected size of amplified wild-type sequence is 935 bp. Arrowheads indicate the isolates with a majority of SU (surface subunit) PCR products shorter than consensus. All the other PCR amplifications (set 1 for other isolates and all of set 2) not shown here gave products of expected sizes. (B–D) Partial electropherograms showing the sequence reading spanning the deleted genomic regions in the three isolates arrowheaded in (A). The start and end sites of the deleted region (382 nt) in (B) are within nt4860–4869 and nt5242–5251, respectively, but are uncertain due to identical sequences flanking the deletion. Numbers above each electropherogram indicate the nucleotide positions according to the reference BLV *env* sequence (K02120). Deduced amino acid sequences and termination sites are shown at the bottom. Bolded amino acids, corresponding to underlined nucleotides, indicate where the deduced sequences are in frame with consensus.

infected host cells could lead to the escape from host immune surveillance against the replicating viruses, a strategy widely used in other retroviruses such as HIV-1 (Liu et al., 2006). Such *env*-defective viruses would also be invisible in *Env*-targeting serological diagnostic tests. They could be even undetectable in PCR-based diagnostic screening if one or both of the paired PCR primers were located within the deleted genomic region. For example, the commonly used screening primer set enclosing *env* segment nt5099–5542 (Fechner et al., 1997; Licursi et al., 2002; Zaghawa et al., 2002; Asfaw et al., 2005; Felmer et al., 2005) would not detect any of the three defective forms reported here.

The coding sequence at the beginning of BLV *env* gene is shared by several other genes (Fig. 1): *rex* and *R3* (sharing with *env* in nt4821–4871 in the same reading frame), *pol* (sharing nt4821–4875, second reading frame), and *tax* (sharing nt4868–4872, third reading frame). The deletions in JP AI-1 and JP EH-1 occurred within this overlapping coding region, resulting in abolished wild type *pol* stop codon, deleted *rex* and *R3* original splicing donor sites, and missed initiating coding sequence for *tax* (Table 4). In comparison, the deletion in isolate JP EH-2 (nt4960–5504) reserved intact *pol*, *rex*, *R3*, and *tax* coding frames, although frame-shifting for *env* (Fig. 5D; Table 4). Each of the three bovine isolates containing *env*-deletion mutants contained full-length LTR sequence (Zhao et al., 2007a) and pXBL coding region (Zhao et al., 2007b). We were unable to test any correlation between the *env* deletion mutants and host health status in this study. None of our donor cattle were reported to have malignancies at the time of sample collection. We have peripheral leukocyte counts only for the Japanese cattle, and the counts of the three with *env* deletions did not differ significantly from those of cattle with no *env* deletions (Table 4 and data not shown).

In summary, we added 28 new sequences to the current literature of 16 full-length BLV *env* sequences, and presented the most comprehensive population study to date for full-length BLV *env* gene sequences. Analyses on BLV *env*-based clustering (phylogeny, genotyping), sequence polymorphisms, and possible selection patterns expand available literature and provide information to help resolve certain controversial issues (e.g., relative variability in different BLV *env* sub-coding regions). The massive natural deletions observed in this study might indicate the possible escape strategy of transformation-competent, replication-defective BLV proviral species. Hopefully, our work would serve as the base for further functional studies on deltaretroviral *envelope* genes/proteins.

Materials and methods

Cell culture and biological samples

BLV-producing cell lines FLK (fetal lamb kidney) (Van Der Maaten and Miller, 1975) and Bat_2Cl_6 (bat lung epithelial) (Graves and Ferrer, 1976) were maintained in Dulbecco's modified Eagle's medium (Life Technologies, Inc.) with 2 mM L-glutamine, 10 μg/ml insulin, 100 μg/ml streptomycin (Pfizer), 100 U/ml penicillin, 50 U/ml polymyxin B and 10% fetal bovine serum (FBS) (all from Sigma Chemical Co.). Cells were grown at 37 °C in a moist atmosphere of 5% CO_2 and passaged at confluence. DNA samples US IA (from FLK) and US PA (from Bat_2Cl_6) were extracted from cell lines, and the other US (United States) and CR (Costa Rican) samples were genomic DNA extracted from the peripheral blood leukocytes of BLV-infected dairy cattle, using Qiagen DNeasy Tissue Kit (Qiagen). Genomic DNA samples (containing proviral BLV DNA) BE and JP (Japanese) series were generous gifts from other researchers, obtained as already-extracted DNA. Standard precautions were used to prevent molecular cross-contamination, including no DNA extraction or preparation of reaction mixture in the laboratory used for post-PCR amplification work.

Polymerase chain reaction (PCR)

Nested PCR was used to amplify segments of the BLV *env* region. All reagents except primers were from Promega (Madison, WI, USA). The final reaction mixture (50 μl/sample) contained 1× buffer, 2.0 mM $MgCl_2$, 200 μM dNTPs, 200 μM of each primer and 1.5 U *Taq* polymerase. Sterile hot start reaction tubes (Molecular Bio Products) were used to reduce non-specific binding. Primers (synthesized by Operon Biotechnologies, Huntsville, AL, USA) (Fig. 1) are: Set 1: Out 5′ (nt4700–4720) CCGGGCAAAACAATCGTCGGT, Out 3′ (nt5810–5791) CTGATGGCTAAGGGCAGACA, In 5′ (nt4763–4778) CTCTCCTGGCTACTGA, In 3′ (nt5697–5683) AGGAAGCCGTAGAGA; Set 2: Out 5′ (nt5600–5617) ATGGCACCACCCTTCCCA, Out 3′ (nt6470–6454) GAACTCGGTGCTGGCAG, In 5′ (nt5638–5651) CTCAAGGCAACGCC, In 3′ (nt6429–6414) CTGACTACTCCACTAA. Primers labeled "Out" were used for the first round of amplification, and those labeled "In" for the second round of amplification and lie within the sequence amplified by the "Out" primers. The cycle conditions were: 94 °C, 3 min; 30 (Out) or 42 (In) cycles of: 94 °C, 30 s; 54 °C (Out) or 49 °C (In), 45 s; 68 °C, 30 s; then 72 °C, 7 min. After amplification of 50 μl reaction, 3 μl of product was checked by electrophoresis on a 1% agarose gel stained with 1% ethidium bromide and viewed under UV light. The Zymo Spin PCR Purification kit (Zymo Research Cooperation) was used to purify the products for sequencing. At least three independent PCR products of each amplified *env* segment were sequenced at the University of California Sequencing Facility (Berkeley). The mixture sent for direct sequencing was 6 μl of purified PCR product mixed with 0.5 μl of 5 μM sequencing primer: Set 1: 5′ (nt5144–5161) TGATTGCGAGCCCCGATG, 3′ (nt5443–5425) GCCCGTGCTGTTTGATTTA; Set 2: 5′ (nt6013–6032) ATGACTCCATTATCCTCCGC, 3′ (nt6150–6125) GGAACAGGCTTAGAATGAATG. Sequencing was conducted double-directionally, and all sequences were also double-checked against their corresponding electropherograms. For the sequence variations, only those consistently shown in the sequences of all three replicates were accepted as real variations rather than PCR errors. Multiple sequences were aligned using the Gene Inspector Program (Textco BioSoftware, Inc. West Lebanon, New

Hampshire, USA). A consensus sequence was calculated based on the individual set of input sequences.

Phylogenetic analyses, genotyping, and statistics

Phylogenetic trees were constructed using the Neighbor-Joining method and the Minimum-Evolution method in MEGA 3.1 (Kumar et al., 2004), each with 1000 bootstrapping replications. Genotyping based on deduced restriction fragment length polymorphism (RFLP) using 444-bp partial BLV *env* sequences was as described previously (Fechner et al., 1997; Licursi et al., 2002). Statistic analyses were performed using GraphPad InStat 3.06 (GraphPad Software, 2003). For comparison of two paired, nonparametric groups of data points (e.g. P_N vs. P_S in gp51), Wilcoxon matched pairs test was used. For multiple comparison (e.g. variability in leader peptide vs. gp51 vs. gp30), Friedman's test with Dunn's post test was used. Synonymous and non-synonymous sites and substitutions were determined using DnaSP 4.10 (Rozas et al., 2003).

Homology modeling and display

The conformation of the BLV gp51 N-terminus (aa34–185) was predicted by homology alignment and modeling based on the crystal structure of the receptor binding domain of ecotropic Friend murine leukemia virus (F-MuLV) surface protein (Fass et al., 1997; Molecular Modeling DataBase [MMDB] ID No. 22594), as described previously (Gatot et al., 2002; Johnston et al., 2002). The conformation of BLV gp30 external domain (aa327–415) was predicted by homology alignment and modeling based on the crystal structure of HTLV-1 gp21 ectodomain (Kobe et al., 1999; MMDB ID No. 10056). The aligned sequences were submitted to ESyPred3D Web Server (http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/) for homology modeling. The structural model PDB files were viewed and annotated using DeepView Swiss-Pdb Viewer (http://www.expasy.org/spdbv/).

Acknowledgments

We thank Dr. Hiroshi Sentsui for tremendous help in field sample and data collection; Drs. Yoko Aida, Bruce Anderson, Dagmar Beier, Yeon Choi, Joy Gardner, David Hayes, and Carlos Jimenez-Sanchez for generous gifts of BLV isolates; Drs. Gary Firestone, Malcolm Potts, and anonymous reviewers for helpful advice; Drs. Cong Liu, Xi Chen, Yang Yang, Hua-Min Shen, Kathleen McGirr, and Jianqiao Xiao for technical support. Authors also thank UC Berkeley DNA Sequencing Facility, especially Dr. Hitomi Asahara and Brian McCarthy, for continuous help. This research was supported in part by funds from the California Breast Cancer Research Program of the University of California (Grant 6PB-0075), the U.S. Army Medical Research and Materiel Command Breast Cancer Research Program (Grant DAMD17-02-1-0320), a Fulbright Scholar Award to G.C. Buehring, and a Sigma Xi Grant-in-Aid of Research and Doctoral Student Fellowships from the Graduate Program in Endocrinology, UC Berkeley, to X. Zhao.

References

Asfaw, Y., Tsuduku, S., Konishi, M., Murakami, K., Tsuboi, T., Wu, D., Sentsui, H., 2005. Distribution and superinfection of bovine leukemia virus genotypes in Japan. Arch. Virol. 150, 493–505.

Ban, J., Czene, S., Altaner, C., Callebaut, I., Krchnak, V., Merza, M., Burny, A., Kettmann, R., Portetelle, D., 1992. Mapping of sequential epitopes recognized by monoclonal antibodies on the bovine leukaemia virus external glycoproteins expressed in *Escherichia coli* by means of antipeptide antibodies. J. Gen. Virol. 73, 2457–2461.

Blanchard, S., Astier-Gin, T., Tallet, B., Moynet, D., Londos-Gagliardi, D., Guillemain, B., 1999. Amino acid changes at positions 173 and 187 in the human T-cell leukemia virus type 1 surface glycoprotein induce specific neutralizing antibodies. J. Virol. 73, 9369–9376.

Bruck, C., Mathot, S., Portetelle, D., Berte, C., Franssen, J.D., Herion, P., Burny, A., 1982. Monoclonal antibodies define eight independent antigenic regions on the bovine leukemia virus (BLV) envelope glycoprotein gp51. Virology 122, 342–352.

Bruck, C., Rensonnet, N., Portetelle, D., Cleuter, Y., Mammerickx, M., Burny, A., Mamoun, R., Guillemain, B., van der Maaten, M., Ghysdael, J., 1984. Biologically active epitopes of bovine leukemia virus glycoprotein gp51: their dependence on protein glycosylation and genetic variability. Virology 136, 20–31.

Buehring, G.C., Philpott, S.M., Choi, K.Y., 2003. Humans have antibodies reactive with bovine leukemia virus. AIDS Res. Hum. Retroviruses 19, 1105–1113.

Callebaut, I., Burny, A., Krchnak, V., Gras-Masse, H., Wathelet, B., Portetelle, D., 1991. Use of synthetic peptides to map sequential epitopes recognized by monoclonal antibodies on the bovine leukemia virus external glycoprotein. Virology 185, 48–55.

Callebaut, I., Voneche, V., Mager, A., Fumiere, O., Krchnak, V., Merza, M., Zavada, J., Mammerickx, M., Burny, A., Portetelle, D., 1993. Mapping of B-neutralizing and T-helper epitopes on the bovine leukemia virus external glycoprotein gp51. J. Virol. 67, 5321–5327.

Camargos, M.F., Stancek, D., Rocha, M.A., Lessa, L.M., Reis, J.K., Leite, R.C., 2002. Partial sequencing of *env* gene of bovine leukaemia virus from Brazilian samples and phylogenetic analysis. J. Vet. Med., B 49, 325–331.

Camargos, M.F., Pereda, A., Stancek, D., Rocha, M.A., Reis, J.K., Greiser-Wilke, I., Leite, R.C., 2006. Molecular characterization of the *env* gene from Brazilian field isolates of bovine leukemia virus. Virus Genes 34, 343–350.

Cantor, G.H., 1996. A potential proline-rich motif upstream of the immunoreceptor tyrosine-based activation motif in bovine leukemia virus gp30, Epstein–Barr virus LMP2A, herpesvirus papio LMP2A, and African horsesickness virus VP7. Virology 220, 265–266.

Chackerian, B., Rudensey, L.M., Overbaugh, J., 1997. Specific N-linked and O-linked glycosylation modifications in the envelope V1 domain of simian immunodeficiency virus variants that evolve in the host alter recognition by neutralizing antibodies. J. Virol. 71, 7719–7727.

Cheng-Mayer, C., Brown, A., Harouse, J., Luciw, P.A., Mayer, A.J., 1999. Selection for neutralization resistance of the simian/human immunodeficiency virus SHIVSF33A variant in vivo by virtue of sequence changes in the extracellular envelope glycoprotein that modify N-linked glycosylation. J. Virol. 73, 5294–5300.

Daniel, R.C., Gatei, M.H., Good, M.F., Boyle, D.B., Lavin, M.F., 1993. Recombinant viral vaccines for enzootic bovine leucosis. Immunol. Cell. Biol. 71, 399–404.

Dimitrov, A.S., Rawat, S.S., Jiang, S., Blumenthal, R., 2003. Role of the fusion peptide and membrane-proximal domain in HIV-1 envelope glycoprotein-mediated membrane fusion. Biochemistry 42, 14150–14158.

Dube, S., Bachman, S., Spicer, T., Love, J., Choi, D., Esteban, E., Ferrer, J.F., Poiesz, B.J., 1997. Degenerate and specific PCR assays for the detection of bovine leukaemia virus and primate T cell leukaemia/lymphoma virus pol DNA and RNA: phylogenetic comparisons of amplified sequences from cattle and primates from around the world. J. Gen. Virol. 78, 1389–1398.

Dube, S., Dolcini, G., Abbott, L., Mehta, S., Dube, D., Gutierrez, S., Ceriani, C., Esteban, E., Ferrer, J., Poiesz, B., 2000. The complete genomic sequence of a BLV strain from a Holstein cow from Argentina. Virology 277, 379–386.

Fass, D., Davey, R.A., Hamson, C.A., Kim, P.S., Cunningham, J.M., Berger, J.M., 1997. Structure of a murine leukemia virus receptor-binding glycoprotein at 2.0 angstrom resolution. Science 277, 1662–1666.

Fechner, H., Blankenstein, P., Looman, A.C., Elwert, J., Geue, L., Albrecht, C., Kurg, A., Beier, D., Marquardt, O., Ebner, D., 1997. Provirus variants of the bovine leukemia virus and their relation to the serological status of naturally infected cattle. Virology 237, 261–269.

Felmer, R., Munoz, G., Zuniga, J., Recabal, M., 2005. Molecular analysis of a 444 bp fragment of the bovine leukaemia virus gp51 *env* gene reveals a high frequency of non-silent point mutations and suggests the presence of two subgroups of BLV in Chile. Vet. Microbiol. 108, 39–47.

Gallaher, W.R., Ball, J.M., Garry, R.F., Martin-Amedee, A.M., Montelaro, R.C., 1995. A general model for the surface glycoproteins of HIV and other retroviruses. AIDS Res. Hum. Retroviruses 11, 191–202.

Gatot, J.S., Callebaut, I., Mornon, J.P., Portetelle, D., Burny, A., Kerkhofs, P., Kettmann, R., Willems, L., 1998. Conservative mutations in the immunosuppressive region of the bovine leukemia virus transmembrane protein affect fusion but not infectivity in vivo. J. Biol. Chem. 273, 12870–12880.

Gatot, J.S., Callebaut, I., van Lint, C., Demonte, D., Kerkhofs, P., Portetelle, D., Burny, A., Willems, L., Kettmann, R., 2002. Bovine leukemia virus SU protein interacts with zinc, and mutations within two interacting regions differently affect viral fusion and infectivity in vivo. J. Virol. 76, 7956–7967.

Goodenow, M.M., Hayward, W.S., 1987. 5' long terminal repeats of myc-associated proviruses appear structurally intact but are functionally impaired in tumors induced by avian leukosis viruses. J. Virol. 61, 2489–2498.

Graves, D.C., Ferrer, J.F., 1976. In vitro transmission and propagation of the bovine leukemia virus in monolayer cell cultures. Cancer Res. 36, 4152–4159.

Hadlock, K.G., Goh, C.J., Bradshaw, P.A., Perkins, S., Lo, J., Kaplan, J.E., Khabbaz, R., Foung, S.K., 1995. Delineation of an immunodominant and human T-cell lymphotropic virus (HTLV)-specific epitope within the HTLV-I transmembrane glycoprotein. Blood 86, 1392–1399.

Hall, W.W., Liu, C.R., Schneewind, O., Takahashi, H., Kaplan, M.H., Roupe, G., Vahlne, A., 1991. Deleted HTLV-I provirus in blood and cutaneous lesions of patients with mycosis fungoides. Science 253, 317–320.

Inabe, K., Nishizawa, M., Tajima, S., Ikuta, K., Aida, Y., 1999. The YXXL sequences of a transmembrane protein of bovine leukemia virus are required for viral entry and incorporation of viral envelope protein into virions. J. Virol. 73, 1293–1301.

Johnson, W.E., Sanford, H., Schwall, L., Burton, D.R., Parren, P.W., Robinson, J.E., Desrosiers, R.C., 2003. Assorted mutations in the *envelope* gene of simian immunodeficiency virus lead to loss of neutralization resistance against antibodies representing a broad spectrum of specificities. J. Virol. 77, 9993–10003.

Johnston, E.R., Radke, K., 2000. The SU and TM envelope protein subunits of bovine leukemia virus are linked by disulfide bonds, both in cells and in virions. J. Virol. 74, 2930–2935.

Johnston, E.R., Albritton, L.M., Radke, K., 2002. Envelope proteins containing single amino acid substitutions support a structural model of the receptor-binding domain of bovine leukemia virus surface protein. J. Virol. 76, 10861–10872.

Kobe, B., Center, R.J., Kemp, B.E., Poumbourios, P., 1999. Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. Proc. Natl. Acad. Sci. U.S.A. 96, 4319–4324.

Kojima, K., Sawada, T., Yasukawa, M., Matsuo, Y., Yakushijin, Y., Narumi, H., Azuma, T., Takimoto, H., Hara, M., 1998. Deleted HTLV provirus in peripheral blood cells of a patient with T-cell prolymphocytic leukaemia. Br. J. Haematol. 100, 567–570.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5, 150–163.

Licursi, M., Inoshima, Y., Wu, D., Yokoyama, T., Gonzalez, E.T., Sentsui, H., 2002. Genetic heterogeneity among bovine leukemia virus genotypes and its relation to humoral responses in hosts. Virus Res. 86, 101–110.

Liu, Y., McNevin, J., Cao, J., Zhao, H., Genowati, I., Wong, K., McLaughlin, S., McSweyn, M.D., Diem, K., Stevens, C.E., Maenza, J., He, H., Nickle, D.C., Shriner, D., Holte, S.E., Collier, A.C., Corey, L., McElrath, M.J., Mullins, J.I., 2006. Selection on the human immunodeficiency virus type 1 proteome following primary infection. J. Virol. 80, 9519–9529.

Mamoun, R.Z., Morisson, M., Rebeyrotte, N., Busetta, B., Couez, D., Kettmann, R., Hospital, M., Guillemain, B., 1990. Sequence variability of bovine leukemia virus *env* gene and its relevance to the structure and antigenicity of the glycoproteins. J. Virol. 64, 4180–4188.

Molteni, E., Agresti, A., Meneveri, R., Marozzi, A., Malcovati, M., Bonizzi, L., Poli, G., Ginelli, E., 1996. Molecular characterization of a variant of proviral bovine leukaemia virus (BLV). Zentralbl. Veterinarmed. Beih. 43, 201–211.

Novakovic, S., Sawai, E.T., Radke, K., 2004. Dileucine and YXXL motifs in the cytoplasmic tail of the bovine leukemia virus transmembrane envelope protein affect protein expression on the cell surface. J. Virol. 78, 8301–8311.

Nunn, M., Chan, S., Duesberg, P.H., 1984. Complete *env* gene deletions of three replication-defective strains of Rous sarcoma virus and a model for the origin of their genetic structures. Virology 134, 466–471.

Ogawa, Y., Sagata, N., Tsuzuku-Kawamura, J., Koyama, H., Onuma, M., Izawa, H., Ikawa, Y., 1987. Structure of a defective provirus of bovine leukemia virus. Microbiol. Immunol. 31, 1009–1015.

Ohshima, K., Kikuchi, M., Yoneda, S., Kobari, S., Sumiyosi, Y., Takeshita, M., Kimura, N., 1993. Restriction of T cell receptor variable region in lymph nodes of adult T cell leukemia/lymphoma. Hematol. Oncol. 11, 147–154.

Penciolelli, J.F., Wendling, F., Robert-Lezenes, J., Barque, J.P., Tambourin, P., Gisselbrecht, S., 1987. Genetic analysis of myeloproliferative leukemia virus, a novel acute leukemogenic replication-defective retrovirus. J. Virol. 61, 579–583.

Portetelle, D., Dandoy, C., Burny, A., Zavada, J., Siakkou, H., Gras-Masse, H., Drobecq, H., Tartar, A., 1989a. Synthetic peptides approach to identification of epitopes on bovine leukemia virus envelope glycoprotein gp51. Virology 169, 34–41.

Portetelle, D., Couez, D., Bruck, C., Kettmann, R., Mammerickx, M., van der Maaten, M., Brasseur, R., Burny, A., 1989b. Antigenic variants of bovine leukemia virus (BLV) are defined by amino acid substitutions in the NH2 part of the envelope glycoprotein gp51. Virology 169, 27–33.

Reeves, R.H., O'Brien, S.J., 1984. Molecular genetic characterization of the RD-114 gene family of endogenous feline retroviral sequences. J. Virol. 52, 164–171.

Reichert, M., Winnicka, A., Willems, L., Kettmann, R., Cantor, G.H., 2001. Role of the proline-rich motif of bovine leukemia virus transmembrane protein gp30 in viral load and pathogenicity in sheep. J. Virol. 75, 8082–8089.

Rice, N.R., Stephens, R.M., Couez, D., Deschamps, J., Kettmann, R., Burny, A., Gilden, R.V., 1984. The nucleotide sequence of the *env* gene and post-*env* region of bovine leukemia virus. Virology 138, 82–93.

Roy-Burman, Y., 2000. A novel truncated *env* gene isolated from a feline leukemia virus-induced thymic lymphosarcoma. J. Virol. 74, 1451–1456.

Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y., Ikawa, Y., 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc. Natl. Acad. Sci. USA 82, 677–681.

Sheets, R.L., Pandey, R., Jen, W.C., Roy-Burman, P., 1993. Recombinant feline leukemia virus genes detected in naturally occurring feline lymphosarcomas. J. Virol. 67, 3118–3125.

Tchenio, T., Heidmann, T., 1991. Defective retroviruses can disperse in the human genome by intracellular transposition. J. Virol. 65, 2113–2118.

Watanabe, N., Yugawa, T., Ikawa, Y., Amanuma, H., 1995. Both the changes of six amino acids and the C-terminal truncation caused by a one-base insertion in the defective *env* gene of Friend spleen focus-forming virus significantly affect the pathogenic activity of the encoded leukemogenic membrane glycoprotein (gp55). J. Virol. 69, 7606–7611.

Westaway, D., Payne, G., Varmus, H.E., 1984. Proviral deletions and oncogene base-substitutions in insertionally mutagenized c-*myc* alleles may contribute to the progression of avian bursal tumors. Proc. Natl. Acad. Sci. U.S.A. 81, 843–847.

Willems, L., Gatot, J.S., Mammerickx, M., Portetelle, D., Burny, A., Kerkhofs, P., Kettmann, R., 1995. The YXXL signalling motifs of the bovine leukemia

virus transmembrane protein are required for in vivo infection and maintenance of high viral loads. J. Virol. 69, 4137–4141.

Willems, L., Burny, A., Collete, D., Dangoisse, O., Dequiedt, F., Gatot, J.S., Kerkhofs, P., Lefebvre, L., Merezak, C., Peremans, T., Portetelle, D., Twizere, J.C., Kettmann, R., 2000. Genetic determinants of bovine leukemia virus pathogenesis. AIDS Res. Hum. Retroviruses 16, 1787–1795.

Zaghawa, A., Beier, D., Abd El-Rahim, I.H., Karim, I., El-ballal, S., Conraths, F.J., Marquardt, O., 2002. An outbreak of enzootic bovine leukosis in upper Egypt: clinical, laboratory and molecular–epidemiological studies. J. Vet. Med., B 49, 123–129.

Zhao, X., Jimenez, C., Sentsui, H., Buehring, G.C., 2007a. Sequence polymorphisms in the long terminal repeat of bovine leukemia virus: evidence for selection pressures in regulatory sequences. Virus Res. 124, 113–124.

Zhao, X., McGirr, K.M., Buehring, G.C., 2007b. Potential evolutionary influences on overlapping reading frames in the bovine leukemia virus pXBL region. Genomics 89, 502–511.